Biometrics 61, 847–855
September 2005

DOI: 10.1111/j.1541-0420.2005.00355.x

Standard Errors for Attributable Risk for Simple and Complex Sample Designs

Barry I. Graubard* **and Thomas R. Fears**

Biostatistics Branch, National Cancer Institute, 6120 Executive Boulevard,
Room 8024, Bethesda, Maryland 20892, U.S.A.
**email:* graubarb@mail.nih.gov

Summary. Adjusted attributable risk (AR) is the proportion of diseased individuals in a population that is due to an exposure. We consider estimates of adjusted AR based on odds ratios from logistic regression to adjust for confounding. Influence function methods used in survey sampling are applied to obtain simple and easily programmable expressions for estimating the variance of $\widehat{AR}$. These variance estimators can be applied to data from case–control, cross-sectional, and cohort studies with or without frequency or individual matching and for sample designs with subject samples that range from simple random samples to (sample) weighted multistage stratified cluster samples like those used in national household surveys. The variance estimation of $\widehat{AR}$ is illustrated with: (i) a weighted stratified multistage clustered cross-sectional study of childhood asthma from the Third National Health and Examination Survey (NHANES III), and (ii) a frequency-matched case–control study of melanoma skin cancer.

Key words: Attributable risk; Influence function; Population attributable fraction; Survey sampling; Taylor deviate.

1. Introduction

The "population attributable risk" (AR) of a disease due to a risk factor is the proportion of diseased individuals in the population that would not develop if the risk factor had been eliminated. The AR of lung cancer due to smoking is the proportion of lung cancer cases that would not have occurred if no one in the population smoked. There are other equivalent terms for AR such as "population attributable fraction" and "population etiologic fraction" (Benichou, 2000, p. 51). The AR can also be defined for a reduction of the level of a risk factor in a population. Methods for estimating unadjusted and (model-based) adjusted ARs for confounders have been developed for case–control (Levin, 1953; Bruzzi et al., 1985), cross-sectional (Basu and Landis, 1993), and cohort (Benichou, 2001) studies. Using the delta method, Benichou and Gail (1990) derived estimates of standard errors for adjusted AR for several case–control study designs. However, their computational formulas are complicated and difficult to implement. Also, it is important to generalize these standard error estimates to surveys with complex sample designs that involve sample weighting and cluster sampling because ARs are estimated from surveys (e.g., Gergen et al., 1998; Gillum, Mussolino, and Madans, 2000), which provide excellent sources for population-based data, and analytical variance estimators are not generally available for complex surveys (Benichou, 2001).

A general approach has been developed in the survey research literature for obtaining estimates of variance of complex estimators under a variety of sample designs that are based on the delta method (Binder, 1996; Deville, 1999; Demnati and Rao, 2001, 2004; Shah, 2002, 2004). For an estimator G, this approach uses results from influence function theory to compute the value of the "Taylor deviate" of the estimator G. A Taylor deviate for G is derived from the first-order Taylor expansion of G for each observation and can be interpreted as a measure of the change (influence) of the value of G when the observation is deleted. It can be shown by delta method arguments that the sum of the Taylor deviates evaluated with the true parameter values approximates G minus its expectation. An estimate of the variance of this sum can be obtained from classical sampling theory for the particular sample design of the study.

In this article, we consider estimates of AR for unmatched, frequency matched, and individual matched case–control, cross-sectional, and cohort studies. Influence function methods are used to obtain Taylor deviates for each of these estimates. Then we provide variance estimates for these AR estimates that are simple to program and are applicable to a wide range of simple and complex sample designs that are used in sample surveys and other observational studies. We show that our variance estimator of AR for unmatched case–control studies is approximately the same as the variance estimator given by Benichou and Gail (1990). We illustrate the estimation of AR and its standard error with a case–control study of melanoma skin cancer and an analysis of environmental smoke using the Third National Health and Nutrition Examination Survey.

2. Definitions of Attributable Risk

Suppose that disease status is indicated by the binary variable $Y = 1$ for disease and $Y = 0$ for no disease, and there is single binary risk factor $E = 1$ for exposed and $E = 0$ for nonexposed. The unadjusted AR of a disease that is due to the risk factor is defined as

$$AR = [\Pr(Y = 1) - \Pr(Y = 1 \mid E = 0)]/\Pr(Y = 1),$$

where Pr is the probability in the population. When there are K levels of the risk factor, $e_0, \ldots, e_{K-1}$ where e_0 is the baseline level, the unadjusted AR for E can be formally defined as

$$AR = [\Pr(Y = 1) - \Pr(Y = 1 \mid E = e_0)] / \Pr(Y = 1).$$

The AR that is adjusted for a confounder with multiple levels, $U = u_1, u_2, \ldots, u_C$, is defined as

$$AR = 1 - \sum_{c=1}^{C} \Pr(U = u_c) \times \Pr(Y = 1 \mid E = e_0, U = u_c)/\Pr(Y = 1) \quad (1)$$

(Whittemore, 1982). This definition can be extended to include U as a vector of q confounders and E as a vector of p risk factors where C in equation (1) becomes the number of combinations of the levels of the q components of U.

A convenient way to express the adjusted AR, which is due to Bruzzi et al. (1985), is

$$AR = 1 - \sum_{c=1}^{C} \sum_{m=0}^{K-1} \rho_{mc} RR_{m0|c}^{-1}, \quad (2)$$

where $\rho_{mc} = \Pr(E = e_m, U = u_c \mid Y = 1)$, the joint distribution of the risk factors and the confounders among the diseased, and $RR_{m0\,|\,c} = \Pr(Y = 1 \mid E = e_m, U = u_c)/\Pr(Y = 1 \mid E = e_0, U = u_c)$, which is the adjusted relative risk of disease for exposure e_m compared to baseline exposure e_0 at level c of confounding. This expression for AR is applicable to several study designs but is particularly useful for population-based case–control studies when the disease is rare because the ρ_{mc} are easily estimated when the cases are a population-based sample, and the $RR_{m0|c}$ are estimated from a logistic regression.

3. Estimation of Adjusted Attributable Risk

In this section, we consider the estimation of adjusted AR as expressed in (2) under several study designs. We assume a logistic regression model for the probability of disease for the jth individual,

$$\ln \frac{\Pr(Y_j = 1 \mid x_j)}{[1 - \Pr(Y_j = 1 \mid x_j)]} = \beta' x_j,$$

where x_j is a (column) vector of covariates, which for notational convenience includes risk factors and confounders as well as possible interactions between them, and β is the vector of regression parameters.

Surveys used to sample subjects for case–control, cross-sectional, or cohort studies may employ simple random samples or subjects may be sampled with different probabilities depending upon their characteristics. For example, in household surveys where one individual is randomly sampled per household, individuals from smaller households will be sampled at a higher rate than individuals from larger households. Approximately unbiased estimates of relative risks (RRs) and ARs can be obtained by weighting observation j by the inverse of its probability of being included in the sample, which is the sample weight w_j. This type of weighted estimation is called Horvitz–Thompson estimation (Horvitz and Thompson, 1951). For case–control studies, a sample-weighted version of the adjusted AR estimator of Bruzzi et al. (1985) is:

$$\widehat{AR}_1 = 1 - \sum_{j=1}^{t} w_j \frac{y_j}{r_j} \Bigg/ \sum_{j=1}^{t} w_j y_j, \quad (3)$$

where t is the total sample size of cases and controls, y_j is the value of Y_j indicating disease status for observation j, r_j is a sample weighted estimate of the adjusted RR of disease for x_j compared with x_{0j}, that is, $r_j = \exp[\hat{\beta}'(x_j - x_{0j})]$ with $\hat{\beta}$ obtained from a sample weighted logistic regression (Korn and Graubard, 1999) and x_{0j} a vector of covariates with the risk factors set at the baseline, $E = e_0$, and any confounders remaining at the same values as they are in x_j.

For simple random samples of cases and controls, the sample weights among the cases are a constant and the sample weights among the controls are another constant. The choice of using weighted or unweighted estimates r_j is a tradeoff between robustness to model misspecification and efficiency of the estimates (Scott and Wild, 1986). If the model is misspecified then weighting by the sample weights will produce regression estimates that are approximately unbiased for the misspecified model of the target population from which the sample was selected. Unweighted estimates may be biased. However, the variance of the weighted estimates will usually be larger than those of the unweighted estimates (Korn and Graubard, 1999, p. 172–176). When the r_j are unweighted then (3) reduces to the unweighted adjusted AR given by Bruzzi et al. (1985).

$\widehat{AR}_1$ can also be used with data from cross-sectional and cohort studies of a rare disease for which the RR can be estimated using a logistic regression. In case of nonrare diseases the adjusted AR as expressed in (1) can be estimated directly as

$$\widehat{AR}_2 = 1 - \sum_{j=1}^{t} w_j p(x_{0j}, \hat{\beta}) \Bigg/ \sum_{j=1}^{t} w_j p(x_j, \hat{\beta}), \quad (4)$$

where $p(x, \hat{\beta}) = \exp(\hat{\beta}' x)/[1 + \exp(\hat{\beta}' x)]$ is the estimated probability of disease from a sample weighted logistic regression for an individual j given x. If the sample weighted logistic regression is used to estimate β then the denominator in $\widehat{AR}_2$ equals the estimated number of cases in the population.

Conditional logistic regression is used to estimate adjusted RRs for case–control studies that individually match controls to cases, for example, sibling controls. Also, conditional logistic analysis can be used to estimate RRs for cross-sectional or cohort studies that sample small clusters of subjects, for example, household or family, and that condition on dummy covariates identifying the subjects from the same cluster. These clusters are treated as matched sets in conditional logistic analyses. Let there be s matched sets with n_k and m_k cases and controls, $k = 1, 2, \ldots, s$. We assume that we have a sample of matched sets where the kth sampled set has a sample

weight of w_k. We do not consider conditional analyses for studies where we sample cases or controls so that sample weights vary among cases and controls within a matched set. For example, for a matched sibling case–control study we would not allow the sibling controls to be sampled at rates that differ by age. The weighted (adjusted) AR estimator can be written as

$$\widehat{AR}_3 = 1 - \sum_{k=1}^{s} w_k \sum_{i=1}^{n_k} \frac{1}{r_{ki}} \Bigg/ \sum_{k=1}^{s} w_k n_k, \tag{5}$$

where r_{ki} is estimated from a sample weighted conditional logistic regression (see Appendix A).

4. Taylor Deviates for AR Estimators

We use the influence function operator, denoted by $\Delta_i(\cdot)$, to provide a simple derivation for the Taylor deviates of the adjusted $\widehat{AR}$'s. Three properties of the operator, which are stated by Shah (2002) and similarly stated by Deville (1999), are used:

I. The delta operator of a sum of a function of m variables $u_j^{(1)},\ldots,u_j^{(m)}$, $S = \sum_{j=1}^{n} f(u_j^{(1)},\ldots,u_j^{(m)})$, is $\Delta_i(S) = f(u_i^{(1)},\ldots,u_i^{(m)})$, for example, if $S = \sum_{j=1}^{n} y_j$, is $\Delta_i(S) = y_i$.

II. Let $F(\hat{\theta})$ be a differentiable function of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M)'$. Then the delta operator of $F(\hat{\theta})$ is $\Delta_i[F(\hat{\theta})] = \sum_{m=1}^{M} \frac{\partial F}{\partial \theta_m} \Delta_i(\hat{\theta}_m)$. If $\hat{\theta}$ is not a closed form expression, for example, $\hat{\beta}$ from a logistic regression, then evaluating $\Delta_i(\hat{\theta})$ may involve using its estimating equations.

III. The delta operator applied to a sum of a function of observations and parameter estimates, that is, $\sum_{j=1}^{n} H(y_j^{(1)},\ldots,y_j^{(m)},\hat{\theta})$, is evaluated as a combination of **I** and **II**:

$$\Delta_i\left[\sum_{j=1}^{n} H\left(y_j^{(1)},\ldots,y_j^{(m)},\hat{\theta}\right)\right] = H\left(y_i^{(1)},\ldots,y_i^{(m)},\hat{\theta}\right) + \left[\frac{\partial \sum_{j=1}^{n} H\left(y_j^{(1)},\ldots,y_j^{(m)},\hat{\theta}\right)}{\partial \hat{\theta}}\right]' \Delta_i(\hat{\theta}).$$

4.1 *Taylor Deviate for* $\widehat{AR}_1$

The Taylor deviates of $\widehat{AR}_1$, $\widehat{AR}_2$, and $\widehat{AR}_3$ are derived in Appendix A using the three properties of $\Delta_i(\cdot)$. The deviates for $\widehat{AR}_1$ and $\widehat{AR}_2$ are given by:

$$\Delta_i(\widehat{AR}_1) = \frac{-1}{\sum_{j=1}^{t} w_j y_j} \left\{ \frac{w_i y_i}{r_i} - \sum_{j=1}^{t} w_j y_j [(x_j - x_{0j}) r_j^{-1}]' \times \Delta_i(\hat{\beta}) - (1 - \widehat{AR}_1) w_i y_i \right\} \tag{6}$$

and

$$\Delta_i(\widehat{AR}_2) = \frac{-1}{\sum_{j=1}^{t} w_j p(x_j,\hat{\beta})} \Bigg(w_i p(x_{0i},\hat{\beta}) + \sum_{j=1}^{t} w_j x_{0j}' p(x_{0j},\hat{\beta}) \times [1 - p(x_{0j},\hat{\beta})]\, \Delta_i(\hat{\beta}) - \frac{\sum_{j=1}^{t} w_j p(x_{0j},\hat{\beta})}{\sum_{j=1}^{t} w_j p(x_j,\hat{\beta})} \times \Bigg\{ w_i p(x_i,\hat{\beta}) + \sum_{j=1}^{t} w_j x_j' p(x_j,\hat{\beta}) \times [1 - p(x_j,\hat{\beta})]\, \Delta_i(\hat{\beta}) \Bigg\} \Bigg), \tag{7}$$

where $\Delta_i(\hat{\beta}) = \{\sum_{j=1}^{t} w_j x_j x_j' p(x_j,\hat{\beta})[1 - p(x_j,\hat{\beta})]\}^{-1} w_i x_i \times [y_i - p(x_i,\hat{\beta})]$. The expression for $\Delta_i(\hat{\beta})$ is obtained by applying the delta operator to the estimating equations for maximizing the pseudo-likelihood for weighted logistic regression (see Appendix A).

The Taylor deviates of $\widehat{AR}_3$ are given with their derivation in Appendix A. In the computation of $\widehat{AR}_3$, the $\hat{\beta}$ is obtained from maximizing the pseudo-likelihood for a sample weighted conditional logistic regression. In deriving the Taylor deviates, the matched sets are then treated as the fundamental sample units.

5. Variance Estimator for AR Estimates

Let z_i represent a generic Taylor deviate for an estimator of AR, for example, $z_i = \Delta_i(\widehat{AR}_1)$, and let z_i^* be the same as z_i but evaluated at $\hat{\beta} = \beta$. Because $\sum_{i=1}^{t} z_i^*$ for an estimate of AR approximates the estimate minus its expectation, we need only estimate the variance of $\sum_{i=1}^{t} z_i^*$ and then substitute $\hat{\beta}$ for β to obtain estimated variances of the $\widehat{AR}_s$ $s = 1, 2, 3$ (Deville, 1999). We will give a variance estimator for $\sum_{i=1}^{t} z_i^*$ from classical sampling theory when the sample is a stratified multistage cluster sample and then show how this variance estimator can be utilized to obtain variance estimates for a variety of sample designs that are special cases.

In order to introduce stratified multistage cluster sampling and its notation, suppose the population of individuals can be partitioned into a set of primary sampling units (PSUs) and that the PSUs for the population are divided into L sampling strata. For household surveys the PSUs are usually geographically defined, for example, counties, and the strata defined by demographic characteristics, for example, population size of the PSUs. At the first stage of sampling, t_h PSUs are randomly sampled from each stratum h, $h = 1,\ldots,L$. There can be additional stages of sampling nested within the sampled PSUs to obtain a random sample of t_{hi} individuals from the ith sampled PSU in stratum h where each sampled individual has a sample weight w_{hij}, $j = 1,\ldots,t_{hi}$. Stratified multistage cluster sampling is often used in household surveys to select subjects for cross-sectional studies. Also, these surveys are

used to form baseline samples to be followed up in cohort studies (Korn and Graubard, 1999). Stratified multistage cluster sampling can also be used to select controls and/or cases in population-based case–control studies (Graubard, Fears, and Gail, 1989).

Under stratified multistage cluster sampling, $\widehat{AR}_1$ can be expressed as

$$\widehat{AR}_1 = 1 - \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij}\frac{y_{hij}}{r_{hij}} \Bigg/ \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij}y_{hij}, \quad (8)$$

and $\widehat{AR}_2$ can be expressed as

$$\widehat{AR}_2 = 1 - \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij} \times p(x_{0hij}, \hat{\beta}) \Bigg/ \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij}p(x_{hij}, \hat{\beta}). \quad (9)$$

For the individual matched studies we consider stratified multistage cluster sampling of matched sets where each sampled matched set of cases and controls is contained within the clusters sampled in the last stage of sampling, $\widehat{AR}_3$ can be expressed as

$$\widehat{AR}_3 = 1 - \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}}\left(w_{hij}\sum_{k=1}^{n_{hij}}\frac{1}{r_{hijk}}\right) \Bigg/ \sum_{h=1}^{L}\sum_{i=1}^{t_h}\sum_{j=1}^{t_{hi}} w_{hij}n_{hij}, \quad (10)$$

where t_{hi} is the number of matched sets sampled from the ith sampled PSU in stratum h and n_{hij} is the number of cases in the jth sampled matched set from the hith sampled PSU.

The variance estimation for stratified multistage cluster sampling can be simplified if the first stage finite population correction factors are ignored, that is, approximate the first stage sampling of PSUs as sampling with replacement; see the Discussion for the implications of this simplification. In this case, the variance estimator for the AR estimate is expressed in terms of the variability of the between PSU-level sums of the Taylor deviates within the sampling strata. The variance estimate of $\widehat{AR}_s$, $s = 1$, 2, or 3 is given by

$$\widehat{\text{var}}(\widehat{AR}_s) = \sum_{h=1}^{L}\frac{t_h}{t_h - 1}\sum_{i=1}^{t_h}(z_{hi} - \bar{z}_h)^2, \quad (11)$$

where $\bar{z}_h = \frac{1}{t_h}\sum_{i=1}^{t_h} z_{hi}$, $z_{hi} = \sum_{j=1}^{t_{hi}} z_{hij}$, and $z_{hij} = \Delta_{hij}(\widehat{AR}_s)$ (Korn and Graubard, 1999, p. 27–28). Equation (11) is a variance estimator for the sum $\sum_{h=1}^{L}\sum_{i=1}^{t_h} z^*_{hi}$, where z^*_{hi} is the same as z_{hi} but with β substituted for $\hat{\beta}$, that is, z^*_{hi} is estimated by z_{hi}. The indices for the Taylor deviates involve more levels in order to handle the complex sampling, but it should be clear how to relate this notation to that used in the earlier sections of this article. For further details about variance estimation for stratified multistage cluster sampling see Korn and Graubard (1999, p. 19–28).

A point worth noting about the variance estimator in (11) is that besides being applicable to a wide range of study designs that can have complex stratified cluster sampling it is applicable to exposures and covariates that can be continuous as well as categorical. Next, we illustrate the use of expression (11) to obtain variance estimators of AR for two sample designs for sampling cases and controls in case–control studies.

5.1 *Unmatched Case–Control Studies*

We assume that the cases and controls are independent simple random samples from a stratum of cases and a stratum of controls. Because this is a case–control study without individual matching of controls, we will use the estimator $\widehat{AR}_1$. There are two sampling strata ($L = 2$), no cluster sampling ($t_{hi} \equiv 1$), and the sample weights, that is, the inverse of the sampling fraction, are equal to a constant among the cases and a different constant among the controls. Here $\widehat{AR}_1$ reduces to (3) with the weights under the sums canceling out. For this case, the variance estimator reduces to

$$\widehat{\text{var}}(\widehat{AR}_1) = \sum_{h=1}^{2}\frac{t_h}{t_h - 1}\sum_{i=1}^{t_h}(z_{hi} - \bar{z}_h)^2, \quad (12)$$

where

$$z_{hi} = \Delta_{hi}(\widehat{AR}_1) = \frac{-1}{t_1}\left\{\frac{y_{hi}}{r_{hi}} - \sum_{h=1}^{2}\sum_{j=1}^{t_h} y_{hj}\left[(x_{hj} - x_{0hj})r_{hj}^{-1}\right]'\Delta_{hi}(\hat{\beta})\right\}, \quad (13)$$

t_1 is the number of cases and $\bar{z}_h = \frac{1}{t_h}\sum_{i=1}^{t_h} z_{hi}, h = 1, 2$, are the mean of the Taylor deviates for the cases and controls, respectively. Comparing z_{hi} in (13) without the sample weights to (6) the term $(1 - \widehat{AR}_1)\, y_{hi}$ is dropped in (13) because the denominator in $\widehat{AR}_1$, which is the number of cases, is a constant.

5.2 *Frequency Matched Case–Control Studies*

We assume that a stratified simple random sample of cases is selected from L_1 strata where the sampling weights vary by stratum. The controls are frequency matched to the cases within L_2 categories of confounders, for example, the matching could be within gender by 5-year age categories. For purposes of variance estimation, these matching categories are treated as sampling strata for the controls. The matching can be adjusted for in the logistic regression estimation of the RRs by including as covariates dummy variables for the matching categories. Usually the estimates of the RRs from frequency matched case–control studies are not weighted by the sample weights because this would unbalance the matching. Therefore, the adjusted RRs will be unweighted but sample weighting will be used to estimate the risk factor distribution in the case population when estimating the AR. For this sample design the $\widehat{AR}_1$ can be written as

$$\widehat{AR}_1 = 1 - \sum_{h=1}^{L_1}\sum_{i=1}^{t_h} w_h\frac{1}{r_{hi}} \Bigg/ \sum_{h=1}^{L_1}\sum_{i=1}^{t_h} w_h,$$

where the w_h are sample weights for the sampled cases in the L_1 strata. Applying the variance estimator in (11) to this sample design, which has no cluster sampling and $L = L_1 + L_2$ strata for sampling the cases and controls, we obtain

$$\widehat{\text{var}}(\widehat{AR}_1) = \sum_{h=1}^{L} \frac{t_h}{t_h - 1} \sum_{i=1}^{t_h} (z_{hi} - \bar{z}_h)^2,$$

where $\bar{z}_h = \frac{1}{t_h} \sum_{j=1}^{t_h} z_{hi}$ and $z_{hi} = \Delta_{hi}(\widehat{AR}_1)$.

6. Comparison to the Benichou–Gail Variance Estimator

For an unmatched case–control study of a simple random sample of cases and controls, we analytically compared the variance estimator of Benichou and Gail (1990) to our variance estimator given in (12). Benichou and Gail (1990) express their estimator as a sum of three terms:

$$\begin{aligned}\widehat{\text{var}}(\widehat{AR}_1) = &\sum_{mc} \sum_{m'c'} r_{m|c}^{-1} r_{m'|c'}^{-1} \widehat{\text{cov}}\left(\hat{\rho}_{mc}, \hat{\rho}_{m'c'}\right) \\ &+ \sum_{mc} \sum_{m'c'} \hat{\rho}_{mc} \hat{\rho}_{m'c'} \widehat{\text{cov}}\left(r_{m|c}^{-1}, r_{m'|c'}^{-1}\right) \\ &+ \sum_{mc} \sum_{m'c'} r_{m|c}^{-1} \hat{\rho}_{m'c'} \widehat{\text{cov}}\left(\hat{\rho}_{mc}, r_{m'|c'}^{-1}\right), \quad (14)\end{aligned}$$

where $m = 0, \ldots, K-1$ are the levels of exposure and $c = 1, \ldots, C$ are the levels of the confounders. The estimator in (12) can be written as a sum of three terms. Let $z_{hi}^{(1)} = -y_{hi}/t_1 r_{hi}$ and $z_{hi}^{(2)} = \frac{-1}{t_1} \sum_{s=1}^{2} \sum_{j=1}^{t_s} y_{sj} [(x_{sj} - x_{0sj}) r_{sj}^{-1}]' \Delta_{hi}(\hat{\beta})$, then we can re-express (12) as

$$\begin{aligned}\widehat{\text{var}}(\widehat{AR}_1) = &\frac{t_1}{t_1 - 1} \sum_{i=1}^{t_1} \left(z_{1i}^{(1)} - \bar{z}_1^{(1)}\right)^2 \\ &+ \sum_{h=1}^{2} \frac{t_h}{t_h - 1} \sum_{i=1}^{t_h} \left(z_{hi}^{(2)} - \bar{z}_h^{(2)}\right)^2 \\ &+ \frac{2t_1}{t_1 - 1} \sum_{i=1}^{t_1} \left(z_{1i}^{(1)} - \bar{z}_1^{(1)}\right)\left(z_{1i}^{(2)} - \bar{z}_1^{(2)}\right),\end{aligned}$$

which corresponds to the three terms in (14). If we use the robust estimator version of $\widehat{\text{cov}}(r_{m|c}^{-1}, r_{m'|c'}^{-1})$ in (14), as described in Benichou and Gail (1989), and we drop the terms $t_h/(t_h - 1)$ in (11), which are approximately one for this case, then the two estimators are the same. Thus, the simulation results for case–control studies for small samples given in Benichou and Gail (1990) will apply as well to our proposed variance estimator.

7. Two Data Examples

We illustrate the weighted estimation of the adjusted AR and its estimates of variance using data from two studies. The first study is a cross-sectional study of environmental tobacco smoke (ETS) and asthma among children in the Third National Health and Examination Survey (NHANES III). The second study is a population-based case–control study of association sunlight exposure and other risk factors on development of melanoma skin cancer among adults.

7.1 *NHANES III Analysis of ETS and Childhood Asthma*

The NHANES III was conducted from 1988 to 1994 and collected a U.S. national cross-sectional random sample of civilian noninstitutionalized children 2 months to 5 years of age, using a stratified multistage clustered probability sample design. For purposes of variance estimation the sample design is approximated by the sampling of two (pseudo-) PSUs from 49 geographically based (pseudo-) sampling strata (Ezzati et al., 1992; NCHS, 1994). There is a sample weight for each child, which reflects higher probabilities for selecting black American and Mexican–American children and adjustments for differential nonresponse and poststratification to U.S. population sizes.

Following the analysis of Gergen et al. (1998), logistic regression was used to model the prevalence of asthma for 7680 children who were white, black, or Mexican-American and completed the home interview; children of other races were excluded. The NHANES III data and documentation used for this example are available from the United States Centers for Disease Control/National Center for Health Statistics website (`http://www.cdc.gov/nchs/about/major/nhanes/datalink.htm#NHANESIII`). The binary dependent variable, asthma, was a self report (by a parent) of a physician's diagnosis of asthma. ETS, the exposure of interest, was parental self reported usual daily total number of cigarettes smoked by individuals living in the same home as the sampled child. The ETS was categorized into no smoking, 1–19, and $\geq$20 cigarettes. The other confounding independent variables in the regression model were age in months at last birthday, sex, race/ethnicity, birth weight, attendance at day care with $\geq$6 children for at least 10 hours per week for more than 1 month, biologic parent with a history of asthma or hay fever, child was breast fed for at least 1 month, highest grade completed by head of household, and number of persons living in the same household as the child.

Among children 2 months to 5 years of age living in the United States, a sample weighted estimate of 5.8% was reported to be diagnosed with asthma, which shows that childhood asthma is not a particularly rare condition. $\widehat{AR}_2$ in (9) was used to estimate the adjusted AR.

When estimating AR from studies with complex sample designs such as the NHANES III, we recommend using design-based methods that use sample weighted estimation of the AR and that account for stratification and clustering of the sample design in the estimation of the standard error. By doing so, the repeated sampling-based properties of the estimation of the AR for the target population are correctly reflected by the estimator and its standard error. The (sample) weighted AR estimate of childhood asthma due to ETS was 9.9% with a design-based standard error of 6.3%. To demonstrate the effect of the sample weighting on the estimation of the AR, we compared the sample weighted and unweighted estimates in Table 1. The unweighted estimate of AR was 11.3%, which is about 14% larger than the weighted estimate. This was primarily because the unweighted estimates of the adjusted odds ratios (OR) for the ETS were slightly greater than the weighted estimates (unweighted ORs were 1.15 and 1.93 compared to the weighted ORs of 1.01 and 1.87, for exposure to 1–19 and $\geq$20 cigarettes smoked per day compared to nonsmoking households). Table 1 compares our design-based standard errors that use weighted estimation and account for the stratified cluster sampling to standard errors that use (i) unweighted estimation and account for the stratified cluster sampling, (ii) unweighted estimation and do not

Table 1
Adjusted AR and standard error of asthma due to ETS for NHANES III children 2 months to 5 years of age by sample weighting and variance estimation

Sample weighting	$\widehat{AR}$ (%)	Standard error of $\widehat{AR}$ (%)	
		Account for stratified cluster sampling	Ignore stratified cluster sampling
Weighted	9.9	6.3	5.8
Unweighted	11.3	4.8	3.8

account for the stratified cluster sampling, treating the sample as a simple random sample of individual children, or (iii) weighted estimation and do not account for the stratified cluster sampling, treating the sample as a random sample of children with replacement but with differential probabilities of selection. The standard errors that accounted for the clustering were larger because of intracluster correlation of the prevalence of asthma. One can also see that the sample weighting increases the standard errors, which is often the case (Korn and Graubard, 1999, p. 172–176). However, the weighted estimates are approximately unbiased for the target U.S. population.

7.2 *Population-Based Case–Control Study of Risk Factors for Melanoma Skin Cancer*

In a case–control study, all patients aged 20–79 years with histologically confirmed invasive cutaneous melanoma were recruited from those newly diagnosed in 1991–1992 at the University of Pennsylvania's Pigmented Lesion Clinic in Philadelphia and the University of California's Melanoma Clinic in San Francisco. Controls were from outpatient clinics with catchment areas similar to the two melanoma clinics and were frequency matched to patients within strata defined by sex, age group, and study site. Initial complaints of the eligible controls varied widely: about 40% were seen for routine physical examinations, 20% for cardiovascular examination, 10% for infections, and 30% for other reasons. Patients with initial complaints of dermatologic or psychiatric problems were excluded.

Each participant was interviewed in person by trained interviewers to obtain individual characteristics. Each participant was examined and freckling pattern, counts of nevi >2 mm, and dysplastic nevi were recorded. Hair color and complexion were assessed by self report. Examiners (physicians and nurses) were uniformly trained and retrained every 6 months by the same instructor. Dysplastic nevus status for each study subject was confirmed by an expert senior examiner (Tucker et al., 1997).

Unconditional logistic regression with terms for the matching strata was used to examine the risk of melanoma among non-Hispanic white males. Risk factors for skin cancer can depend on gender and we chose to restrict this analysis to males. The analysis was also restricted to non-Hispanic whites because there were few individuals in other ethnic/race groups. Dichotomous risk factors of interest included hair color other than dark brown or black; fair complexion; extensive freckling on at least one area of the body; 50 or more moles not more than 2 mm; and presence of any dysplastic nevi. There were no confounders. Note that this example differs from the previous asthma example in that it has multiple risk factors. While melanoma of the skin is an increasing clinical problem, it is a rare event. The age-adjusted incidence rate for both sexes is 20.1 per 100,000 person years (Ries et al., 2003). The case sample was regarded as a simple random sample of area cases and the control sample was regarded as a stratified sample of controls using the matching strata as the sampling strata. Unweighted logistic regression was used for the $\widehat{AR}_1$. There were 392 cases and 502 controls in this analysis. The estimate of its standard error used Taylor deviates derived in Appendix A in equations (A.1) and (A.2). We obtained $\widehat{AR}_1 = 83.0\%$ with standard error of 3.2%. The formula for the Taylor deviate is easily programmed and a program using SAS Version 8.0 (1999) is provided in Appendix B, which also allows for sample weighting.

8. Discussion

In this article, we have provided simple and easily programmable analytical formulas for computing the standard errors for estimated AR for case–control studies that are frequency or individually matched, cross-sectional studies, and cohort studies. We considered estimating standard errors for sample designs used to select study subjects that ranged from simple random sampling to stratified multistage cluster sampling. Influence function theory in conjunction with classical sampling theory was applied in deriving the formulas for variance estimators of adjusted AR.

Jackknife and bootstrap replication methods for variance estimation offer alternatives to analytical methods for computing standard errors and confidence intervals of AR (Llorca and Delgado-Rodriguez, 2000) but require more computational time. Also, replication methods are readily applicable to complex weighted sample designs (Rust and Rao, 1996). Because of the large sample sizes in our two data examples, the jackknife estimates agree with our estimates to four decimal places as would be expected because for differentiable functions of the data the jackknife is asymptotically equivalent to variance estimation based on the delta method (Krewski and Rao, 1981; Efron and Tibshirani, 1993, Chapter 21). One note of caution is that because replication methods or delta methods for estimating the variance rely on large samples to be consistent either approach could be biased for sparse exposures or in case of small sample size.

Ignoring the finite population correction factors results in variance estimation given by (11) more closely approximates the superpopulation variance. The superpopulation variance is usually the variance of interest because it is the variance of the underlying stochastic process that has given rise to the population from which the sample is selected, and it is this process that most analysts want to make inferences about rather than the finite population (see Graubard and Korn, 2002 for further discussion). However, in general, even after dropping the finite population correction factors, the finite population sampling variance estimators will underestimate the superpopulation variance when the sample fractions are not small and there is nonnegligible between (sampling) strata variability in the AR. Graubard and Korn (2002) provide corrections to the repeated sampling variance estimators that

can be used to estimate the "superpopulation" variance of AR.

AR estimates from case–control studies are used to estimate absolute risk of disease by combining them with age-specific disease incidence rates obtained from a population disease registration system and age-specific mortality rates from other causes of disease (Gail et al., 1989). The influence function methods described in this article can be used to obtain standard errors and confidence intervals for absolute risk. It can be shown that the absolute risk of disease for an individual is a monotone increasing function of $(1 - \widehat{AR})r_0$, where $r_0 = \exp(\hat{\beta}' x_0)$ and x_0 are the vector of the individual's covariates and risk factors with the risk factors set at the baseline levels. Because the age-specific disease incidence and mortality rates are usually based on disease registries that have large samples and that are independent of the studies used to estimate the relative risks, they can be treated as fixed and the influence function method described in this article can be directly applied to obtain standard errors for $(1 - \widehat{AR})r_0$. However, if a cohort study is used to estimate the age-specific disease incidence or mortality rates or the AR then care should be exercised to take into account the variability of these rates in the estimation of the standard errors (Benichou and Gail, 1995). These standard errors can be used to form confidence intervals for estimates of absolute risk of disease.

Finally, for simplicity we have restricted the modeling of the adjusted RR used in $\widehat{AR}_1$ to logistic regression. However, other types of regression modeling could be used to estimate adjusted RRs. For instance, Poisson and proportional hazard regression are commonly used to estimate adjusted relative hazards from cohort studies that, under a rare disease assumption, approximate RRs. Using the influence function method described in this article, Taylor deviates can be obtained for these estimated RRs and then substituted into the formula for the deviate of $\widehat{AR}_1$ which can then be used in (11) to estimate the variance of $\widehat{AR}_1$.

References

Basu, S. and Landis, J. R. (1993). Model-based estimation of population attributable risk under cross-sectional sampling. *American Journal of Epidemiology* **142,** 1338–1343.

Benichou, J. (2000). Attributable risk. In *Encyclopedia of Epidemiologic Methods*, M. H. Gail and J. Benichou (eds), 50–63. Chichester: John Wiley & Sons.

Benichou, J. (2001). A review of adjusted estimates of attributable risk. *Statistical Methods in Medical Research* **10,** 195–216.

Benichou, J. and Gail, M. H. (1989). A delta method for implicitly defined random variables. *American Statistician* **43,** 41–44.

Benichou, J. and Gail, M. H. (1990). Variance calculation and confidence intervals for estimates of attributable risk based on logistic models. *Biometrics* **46,** 991–1003.

Benichou, J. and Gail, M. H. (1995). Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics* **51,** 182–194.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology* **22,** 17–22.

Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A., and Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology* **122,** 904–914.

Demnati, A. and Rao, J. N. K. (2001). *Linearization variance estimators for survey data.* Methodology Branch Working Paper, SSMD-2001-010E, Statistics Canada.

Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology* **30,** 17–26.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* **25,** 193–203.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Ezzati, T. M., Massey, J. T., Waksberg, J., Chu, A., and Maurer, K. R. (1992). Sample design: Third National Health and Nutrition Examination Survey. *Vital and Health Statistics* **2**(113).

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57,** 1198–1206.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast-cancer for white females who are examined annually. *Journal of the National Cancer Institute* **81,** 1879–1886.

Gergen, P. J., Fowler, J. A., Maurer, K. R., Davis, W. W., and Overpeck, M. D. (1998). The burden of environmental tobacco smoke exposure on respiratory health of children 2 months through 5 years of age in the United States: Third National Health and Nutrition Examination Survey 1988 to 1994. *Pediatrics* **101,** e8.

Gillum, R. F., Mussolino, M. E., and Madans, J. H. (2000). Diabetes mellitus, coronary heart disease incidence, and death from all causes in African American and European American women—The NHANES I Epidemiologic Follow-Up Study. *Journal of Clinical Epidemiology* **53,** 511–518.

Graubard, B. I. and Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science* **17,** 73–96.

Graubard, B. I., Fears, T. R., and Gail, M. H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control studies. *Biometrics* **45,** 1053–1071.

Horvitz, D. G. and Thompson, D. J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* **9,** 1010–1019.

Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* **9,** 531–541.

Llorca, J. and Delgado-Rodriguez, M. (2000). A comparison of several procedures to estimate the confidence interval

for attributable risk in case-control studies. *Statistics and Medicine* **19,** 1089–1099.

National Center for Health Statistics. (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988–94. *Vital and Health Statistics* **1**(32).

Ries, L. A. G., Eisner, M. P., Kosary, C. L., Hankey, B. F., Miller, B. A., Clegg, L., and Edwards, B. K. (eds). (2003). *SEER Cancer Statistics Review, 1973–1999.* Bethesda, Maryland: National Cancer Institute.

Rust, K. F. and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5,** 283–310.

SAS Institute, Inc. (1999). *SAS/IML User's Guide, Version 8.* Cary, North Carolina: SAS Institute, Inc.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society, Series B* **48,** 170–182.

Shah, B. V. (2002). Calculus of Taylor deviations. Presented at the Joint Statistical Meetings, New York.

Shah, B. V. (2004). Comment on "Linearization variance estimators for survey data" by A. Demnati and J. N. K. Rao. *Survey Methodology* **30,** 29.

Tucker, M. A., Halpern, A., Holly, E. A., Hartge, P., Elder, D. E., Sagebiel, R. W., Guerry, D., and Clark, W. H. (1997). Clinically recognized dysplastic nevi, a central risk factor for cutaneous melanoma. *Journal of the American Medical Association* **277,** 1439–1444.

Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Statistics and Medicine* **1,** 229–243.

Received November 2003. *Revised November* 2004. *Accepted January* 2005.

Appendix A

Derivation of Taylor Deviates for $\widehat{AR}_1$, $\widehat{AR}_2$, and $\widehat{AR}_3$

A.1 *Taylor Deviate for $\widehat{AR}_1$*

We can write $\widehat{AR}_1 = 1 - \frac{S_1}{S_2}$, where $S_1 = \sum_{j=1}^{t} w_j \frac{y_j}{r_j}$ and $S_2 = \sum_{j=1}^{t} w_j y_j$. The Taylor deviate of $\widehat{AR}_1$ is derived, using the three properties of $\Delta_i(\cdot)$ in Section 4, as

$$\begin{aligned}\Delta_i(\widehat{AR}_1) &= \frac{\partial \widehat{AR}_1}{\partial S_1}\Delta_i(S_1) + \frac{\partial \widehat{AR}_1}{\partial S_2}\Delta_i(S_2)\\ &= \frac{\partial \widehat{AR}_1}{\partial S_1}\left[\frac{w_i y_i}{r_i} + \frac{\partial S_1}{\partial \hat{\beta}}\Delta_i(\hat{\beta})\right] + \frac{\partial \widehat{AR}_1}{\partial S_2} w_i y_i\\ &= \frac{-1}{\sum_{j=1}^{t} w_j y_j}\left[\frac{w_i y_i}{r_i} + \sum_{j=1}^{t} w_j y_j \left(\frac{\partial r_j^{-1}}{\partial \hat{\beta}}\right)' \right.\\ &\qquad \left. \times \Delta_i(\hat{\beta}) - (1 - \widehat{AR}_1) w_i y_i\right], \qquad \text{(A.1)}\end{aligned}$$

where $\partial r_j^{-1}/\partial\hat{\beta} = -(x_j - x_{0j})r_j^{-1}$. The expression for $\Delta_i(\hat{\beta})$ is obtained by applying the delta operator to the estimating equations for maximizing the pseudo-likelihood for sample weighted logistic regression:

$$\begin{aligned}0 &= \Delta_i\left(\sum_{j=1}^{t} w_j x_j [y_j - p(x_j, \hat{\beta})]\right)\\ &= w_i x_i [y_i - p(x_i, \hat{\beta})]\\ &\quad - \left\{\sum_{j=1}^{t} w_j x_j x_j' p(x_j, \hat{\beta})[1 - p(x_j, \hat{\beta})]\right\}\Delta_i(\hat{\beta}).\end{aligned}$$

Solving for $\Delta_i(\hat{\beta})$

$$\begin{aligned}\Delta_i(\hat{\beta}) &= \left\{\sum_{j=1}^{t} w_j x_j x_j' p(x_j, \hat{\beta})[1 - p(x_j, \hat{\beta})]\right\}^{-1}\\ &\quad \times w_i x_i [y_i - p(x_i, \hat{\beta})]. \qquad \text{(A.2)}\end{aligned}$$

A.2 *Taylor Deviate for $\widehat{AR}_2$*

We can write $\widehat{AR}_2 = 1 - (T_1/T_2)$, where $T_1 = \sum_{j=1}^{t} w_j p(x_{0j}, \hat{\beta})$ and $T_2 = \sum_{j=1}^{t} w_j p(x_j, \hat{\beta})$. The Taylor deviate of $\widehat{AR}_2$ given in (7) can be derived in the following way:

$$\begin{aligned}\Delta_i(\widehat{AR}_2) &= \frac{\partial \widehat{AR}_2}{\partial T_1}\Delta_i(T_1) + \frac{\partial \widehat{AR}_2}{\partial T_2}\Delta_i(T_2)\\ &= \frac{\partial \widehat{AR}_2}{\partial T_1}\left[w_i p(x_{0i}, \hat{\beta}) + \frac{\partial T_1}{\partial \hat{\beta}}\right]\\ &\quad + \frac{\partial \widehat{AR}_2}{\partial T_2}\left[w_i p(x_i, \hat{\beta}) + \frac{\partial T_2}{\partial \hat{\beta}}\right]\\ &= \frac{-1}{T_2}\left(w_i p(x_{0i}, \hat{\beta}) + \sum_{j=1}^{t} w_j x_{0j}' p(x_{0j}, \hat{\beta})\right.\\ &\qquad \times [1 - p(x_{0j}, \hat{\beta})]\Delta_i(\hat{\beta})\\ &\qquad - \frac{T_1}{T_2}\left\{w_i p(x_i, \hat{\beta}) + \sum_{j=1}^{t} w_j x_j' p(x_j, \hat{\beta})\right.\\ &\qquad \left.\left. \times [1 - p(x_j, \hat{\beta})]\,\Delta_i(\hat{\beta})\right\}\right). \qquad \text{(A.3)}\end{aligned}$$

A.3 *Taylor Deviate for $\widehat{AR}_3$*

The Taylor deviate of $\widehat{AR}_3$ for a matched set is given by

$$\begin{aligned}\Delta_h(\widehat{AR}_3) &= \frac{-1}{\sum_{k=1}^{s} w_k n_k}\\ &\quad \times \left[w_h \sum_{i=1}^{n_h} \frac{1}{r_{hi}} + \sum_{k=1}^{s} w_k \sum_{i=1}^{n_k}\left(\frac{\partial r_{ki}^{-1}}{\partial \hat{\beta}}\right)'\right.\\ &\qquad \left. \times \Delta_h(\hat{\beta}) - \left(1 - \widehat{AR}_3\right) w_h n_h\right]. \qquad \text{(A.4)}\end{aligned}$$

The derivation of (A.4) follows from applying the properties of $\Delta(\cdot)$ that are given in Section 4. The $\Delta_h(\hat{\beta})$ in (A.4) is

based on the estimating equations from maximizing a pseudo-likelihood for a sample weighted conditional logistic regression. This pseudo-likelihood is given by

$$\ell(\beta) = \prod_{k=1}^{s} \left[\frac{\exp\left(\sum_{i=1}^{t_k} \beta' x_{ki} y_{ki} \right)}{\sum_j \exp\left(\sum_{i_j=1}^{t_k} \beta' x_{ki_j} y_{ki_j} \right)} \right]^{w_k},$$

where x_{ki} are the covariate vectors and y_{ki} are the binary response variables for observation i in matched set k and the summation over j in the denominator is over all combinations of assigning n_k cases and m_k controls among the $t_k = n_k + m_k$ observations in matched set k. The estimating equations for β are obtained by differentiating the log of the pseudo-likelihood with respect to β and are given by

$$0 = U(\hat{\beta}) = \sum_{k=1}^{s} w_k \sum_{i=1}^{t_k} x_{ki} y_{ki} - \sum_{k=1}^{s} w_k \frac{A_k}{B_k},$$

where

$$A_k = \sum_j \left(\sum_{i_j=1}^{t_k} x_{ki_j} y_{ki_j} \right) \exp\left(\sum_{i_j=1}^{t_k} \hat{\beta}' x_{ki_j} y_{ki_j} \right)$$

and

$$B_k = \sum_j \exp\left(\sum_{i_j=1}^{t_k} \hat{\beta}' x_{ki_j} y_{ki_j} \right).$$

Applying the delta operator to the estimating equations for each matched set h we obtain $\Delta_h(\hat{\beta})$ as follows:

$$0 = \Delta_h[U(\hat{\beta})] = w_h \sum_{i=1}^{t_h} x_{hi} y_{hi} - w_h \frac{A_h}{B_h} - \frac{\partial \sum_{k=1}^{s} w_k \frac{A_k}{B_k}}{\partial \hat{\beta}} \Delta_h\left(\hat{\beta}\right).$$

Solving for $\Delta_h(\hat{\beta})$

$$\Delta_h(\hat{\beta}) = \left[\sum_{k=1}^{s} w_k \left(\frac{1}{B_k} \frac{\partial A_k}{\partial \hat{\beta}} - \frac{A_k}{B_k^2} \frac{\partial B_k}{\partial \hat{\beta}} \right) \right]^{-1} \times w_h \left(\sum_{i=1}^{t_h} x_{hi} y_{hi} - \frac{A_h}{B_h} \right),$$

where

$$\frac{\partial A_k}{\partial \hat{\beta}} = \sum_j \left(\sum_{i_j=1}^{t_k} x_{ki_j} y_{ki_j} \right) \left(\sum_{i_j=1}^{t_k} x'_{ki_j} y_{ki_j} \right) \times \exp\left(\sum_{i_j=1}^{t_k} \hat{\beta}' x_{ki_j} y_{ki_j} \right)$$

and

$$\frac{\partial B_k}{\partial \hat{\beta}} = A'_k.$$

This expression for $\Delta_h(\hat{\beta})$ is substituted in (A.4) to obtain the Taylor deviate for $\widehat{AR}_3$. $\Delta_h(\hat{\beta})$ without sample weighting was derived by Fay and Graubard (1999) to form a sandwich variance estimator for $\hat{\beta}$.

Appendix B

SAS Program for Taylor Deviate of $\widehat{AR}_1$

The example code is for SAS Version 8.0 (1999) and uses *proc iml*. Males3 is the data set name, $x00$ has a value of one, $m1, \ldots, m11$ can be strata indicator or confounder variables, $x1, \ldots, x5$ are dichotomous risk factors, p is the predicted probability of case status from a logistic regression model of $m1, \ldots, m11$ and $x1, \ldots, x5$ on case/control status, the variable *case* is the case/control indicator, r is the estimated relative odds, and w is the sample weight.

```
proc iml;
use males3;

read all var {x00 m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11
x1 x2 x3 x4 x5 p case rw};

pq=p#(1 - p);
x=x1||x2||x3||x4||x5;
casecnt = case` * (case#w);
x0=(x00||m1||m2||m3||m4||m5||m6||m7||m8||m9||m10||m11);
rinv = r##-1;
xxpq = (x0||x)` * ((x0||x)#pq#w);
caseterm = case * inv(casecnt)#rinv#w;
casectlterm = inv(casecnt) * (case#rinv#w)` * (x);
partialBwrtWl=((inv(xxpq))*((x0||x)#(case-p)#w)`)
[13:17,];
/*must drop first 12 rows that refer to x00 and the
strata/confounder variables*/
onemAR = (case`*(case#w#rinv))*inv(casecnt);
Arterm = (case*inv(casecnt)*onemAR)#w;
z = -caseterm +(casectlterm*partialBwrtWl)`+
    ARterm;
out = z||case;
/*Taylor deviate of AR1 estimator and case indicator
*/
varname_out = {'AR1_deviate' 'Case'};
create deviates from out[colname = varname_out];

append from out;

quit; run;
```

The output data set *out* contains the variable *AR_deviate*, the Taylor deviate of $\widehat{AR}_1$, and the variable *case*. These data can then be merged with the sample design variables that have codes for the sampling strata and cluster memberships for each observation. This merged data set is used as the input data for equation (11) to obtain the variance estimate for $\widehat{AR}_1$.